PCT/SG2004/000366

PA 1249443

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

November 17, 2004

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: *60/517,928*
FILING DATE: *November 06, 2003*

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

M. Sias

M. SIAS
Certifying Officer

Please type a plus sign (+) inside this box → +

PTO/SB/16 (10-01)
Approved for use through 10/31/2002. OMB 0651-0032
Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a requ st f r filing a PROVISIONAL APPLICATION FOR PATENT und r 37 CFR 1.53 (c).

Expr ss Mail Label No. EL 623308768 US

INVENTOR(S)

Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
Chaoqiang	LIU	Singapore
Tao	XIA	Singapore

□ *Additional inventors are being named on the _____ separately numbered sheets attached hereto*

TITLE OF THE INVENTION (280 characters max)

WAVELET DOCUMENT IMAGE COMPRESSION

CORRESPONDENCE ADDRESS

Direct all correspondence to:

☒ Customer Number 27572

OR *Type Customer Number here*

□ Firm *or* Individual Name	Harness, Dickey & Pierce, P.L.C.				
Address	P.O. Box 828				
Address					
City	Bloomfield Hills	State	MI	ZIP	48098
Country	USA	Telephone	248-641-1600	Fax	248-641-0270

ENCLOSED APPLICATION PARTS *(check all that apply)*

☒ Specification *Number of Pages* 9

☒ Drawing(s) *Number of Sheets* 4

☒ Application Data Sheet. See 37 CFR 1.76

□ CD(s), Number

☒ Other (specify) Return Receipt Postcard

☒ Specification Filed in English

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT *(check one)*

☒ Applicant claims small entity status. See 37 CFR 1.27.

☒ A check or money order is enclosed to cover the filing fees

□ The Commissioner is hereby authorized to charge filing fees or credit any overpayment to Deposit Account Number:

□ Payment by credit card. Form PTO-2038 is attached.

FILING FEE AMOUNT ($) 80.00

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☒ No.

□ Yes, the name of the U.S. Government agency and the Government contract number are: _____.

Respectfully submitted,

SIGNATURE

Date 11/6/03

TYPED or PRINTED NAME Paul A. Keller

REGISTRATION NO. *(if appropriate)* 29,752

TELEPHONE (248) 641-1600

Docket Number: 2500-000015

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

Wavelet Document Image Compression

Field Of the Invention

The present invention relates to a novel image compression technique for classifying, matching and identifying document images based on wavelet compression method. This technique is called wavelet document image compression (WDIC) technique. More specifically, WDIC technique relates to separate the character/lines and pictures from the backgrounds of an original document images and to use different techniques to compress each of those components. More generally, this technique may also be applied to other special documents such as particularly important historical documents, scientific papers with mathematical or chemical formulae, software documents and some handwritten signatures.

Background Of the Invention

As electronic storage, retrieval and distribution of documents becomes faster and cheaper, a lot of documents are becoming increasingly digitally. In the last decade existing documents are usually re-typed and converted to HTML or Adobe's PDF format, sometimes are used by Optical Character Recognition (OCR) technique. Unfortunately, these techniques are still far from being able to translate faithfully a scanned document into web page, much of the visual aspect of the original document is likely to be lost. Recently, several authors [1-4] have proposed image-based approaches to digital documents. The "image-based approach" to digital documents is to store and to transmit documents as image. Traditional image compression standards such as JPEG and GIF are inappropriate for document image. Although they are suitable for continuous-tone image (i.e. for most pictures of natural scenes, they are not for the sharp edges of character images. In the other hand, a scan document tends to be quite large if one wants to preserve the readability of the text. It is needed to develop an approach for compression document images that makes it possible to transfer a high-quality of one page of document image at very high compression ratio, the WDIC document image compression technique described here is designed to overcome all the above problems.

Objective Of the Invention

The object of the invention is to provide a novel image compression technique (WDIC) for classifying, matching and identifying document images. A more specific object is to provide a wavelet-based compression algorithm for picture images, and a novel extent-based morphological matching, clustering and wavelet compression algorithm for mostly small character/lines images.

Summary Of the Invention

The invention comprises a number of novel algorithms for an improved document image compression technique. The main idea of our document image compression technique is to extract two main categories of picture areas and character/line areas from the document image and encode the residue image by subtracted these two categories areas from document images.

The character image can be encoded with a novel extent-based morphological matching, clustering and wavelet compression algorithm. A picture image can be encoded with a wavelet-based compression algorithm, which is suitable for grey scale images. The background image also can be encoded with a wavelet-based SAQ compression algorithm.

WDIC is a progressive code. It provides progressive decoding not only on background, but also on character images.

Detailed Description Of the Invention

In the following sections the novel techniques of the WDIC are described. The features of WDIC comprise special image segmentation for a document image, fast classification, morphological matching and clustering algorithm for character/lines images, a wavelet-based compression algorithm for picture images. Results from an actual system of the WDIC showed that the novel means contribute significant performance improvement to two aspects of: highly efficient compression format and a progressive range of compression rate scalability.

Encoder 10

Figure 1 is the block diagram of the encoder **10**. We start our process from a scanned grey scale document image **101** with scanned resolution *r* dpi. Process **102** extracts picture image blocks **103** from **101**. **103** will be encoded by wavelet based SAQ encoder **104** ([5],[6]). Encoder **104** passes the compressed bit stream **105** to the process **118**.

By subtracting the picture images **103** from **101**, the residue image **106** is further processed by **107**. Process **107** generates the connective blocks from residue image **106** based on region growing algorithm.

At first, we posterize the image into 3 levels as below.

$$F(v)=\begin{cases}0 & \text{when } I(v)\geq P\\ 1 & \text{when } 128\leq I(v)<P, \text{ where } I(v) \text{ is the intensity of the pixel at } v=(v_x,v_y).\\ 2 & \text{when } I(v)<128\end{cases}$$

The following algorithm below performs at all untraced pixels u with $F(u)=2$.

1. $S=\phi, S_1=\{u\}$, $W=\frac{r}{72}\times C$.
   C is slightly larger than font size of most characters/letter. (default $C=24$)
2. Find $v\in S_1$, $\{v_i\}_{i=1}^{8}$ represent eight neighbor pixels of v in clockwise order, among them $\{v_1,v_3,v_5,v_7\}$ are 4-neighbor pixels of v. Define $v_{i+8k}=v_i, k\in Z$. $S=S\cup\{v\}, S_1=S_1\setminus\{v\}$

```
3. for $v_i, i = 1,\cdots,8$, $|v_{i,x} - u_x| \le W$, and $|v_{i,y} - u_y| \le W$
    a. if $i = 1,3,5,7$,
        i. if $F(v) = 2$ and $(F(v_i) = 2$ or $(F(v_i) = 1$ and $F(v_{i-1}) + F(v_{i+1}) \ge 1))$
           then $S_1 = S_1 \cup \{v_i\}$
        ii. if $F(v) = 1$ and $(F(v_i) = 2$ and $(F(v_{i-2}) + F(v_{i+2}) \ge 2))$
           then $S_1 = S_1 \cup \{v_i\}$
    b. if $i = 2,4,6,8$,
       if $F(v) = 2$ and $(F(v_i) = 2$ and $F(v_{i-1}) + F(v_{i+1}) \ge 1)$
       then $S_1 = S_1 \cup \{v_i\}$
4. if $S_1 \ne \phi$, go to step 2
5. $A = \{(x,y) \mid x_{min} \le x \le x_{max}, y_{min} \le y \le y_{max}\}$ is character image block.
   where $x_{min} = \min_{v \in S}\{v_x\}, x_{max} = \max_{v \in S}\{v_x\}, y_{min} = \min_{v \in S}\{v_y\}, y_{max} = \max_{v \in S}\{v_y\}$
```

After character image block A is extracted and saved into the character block list. We mark the pixels in this block as the traced pixels and change their value to.255. And same procedure starts from untraced pixels satisfying $F(u)=2$ until no such pixel exists.

Character images **108** are the blocks representing the lines and characters extracted by **107** from **106**. Process **109** clusters the character images hierarchically. We will elaborate the process **109** in **401** to **413**. Process **109** outputs data **110** comprising the character template library and the code of the every character blocks outputted from **109**. The code of character blocks includes the absolute coordinates of the block in the original image and the index of the template it uses. Character encoder **111** encodes the character codes of the character blocks and character template library by SAQ encoder.

The output **112** is compressed bit stream for the characters. Whistle the data **112** is passed to the process **118**, it will be decoded by the decoder **113** which is the counterpart of SAQ encoder **111**. The reconstructed character images **114** are used to get the background image **115**. Process **116** is encoder of wavelet based SAQ encoder for grey scale image ([5], [6]). The compressed bit stream **117** for background image is passed to the process **118**. The process **118** organizes the compressed bit stream of picture image blocks, character image blocks and the background image to generate the compressed data **119** for the whole document image.

Data **119** is organized as the following. We save the document image header and character codes of character blocks and location information of picture image blocks first; then the compressed bit planes corresponding value greater than 2^7 of character template library, picture image blocks and background image will be stored; finally the residue bit plane information will be added one bit plane followed by another from the most significant bit plane to the least significant bit plane. Such organization guarantees the progress decoding of the document image. In the other words, we can obtain the document image from blur version to the finest one.

Picture image block extractor 102

Picture image block extractor 102 is elaborated in the following.

Process 301 estimate the peak value P_0 of histogram of document image, threshold $P = (128 + P_0)/2$, the pixels of intensity of pixel less than P are classified as foreground pixel, other pixels are background pixel.

Process 302 partitions entire document image into blocks with size $W \times W$ where $W = 2^{\lfloor \log_2 r/4 \rfloor}$ and r is the scanned resolution.

Process 303 classify blocks to two types: picture block marked by 1 and nonpicture block marked it by 0. The verdict is based on the statistical features of wavelet decomposition of blocks. The procedure is as following.

Using the wavelet filter to decompose the block once as conventional wavelet decomposition of image. For the computation efficiency, the sum of filter coefficients is 2 and the suggested filter for this procedure is Haar wavelet. The figure below shows this procedure. LL, LH, HL and HH are the notations of lowest frequency component to highest frequency component as usual ([7]).

LL	LH
HL	HH

In generally, a document image is typically composed of a large portion of characters and edges regions, together with a rather small portion of homogeneous regions. Homogeneous regions have the least variation. Characters regions have moderate variation; and lines show the most variation.

$g(c) = \begin{cases} 1 & when\, |c| > A \\ 0 & otherwise \end{cases}$ where A *is* a predefined threshold (default A=16) and c is the wavelet coefficients.

Calculation the sums of wavelet coefficients. The statistical variable we used in classification is following, $count_H = \frac{\sum_{(i,j)\in H} g(C_{i,j})}{1.5W}, where H = HL \cup LH \cup HH$

$average_{LL} = \frac{\sum_{(i,j)\in LL} C_{i,j}}{4S_{LL}}$ where S_{LL} is the total number of wavelet coefficients of LL.

If $count_H < B\ and\ average_{LL} < (P+128)/2$, where B is the predetermined threshold whose default value is 3, the block is marked as picture block, otherwise it is marked as nonpicture block.

Switch **304** checks whether untraced picture block exists. If the answer is **NO**, all picture blocks are saved in data **316** already and finish process **102**.

Otherwise, the next untraced picture block is identified in step **305** and change its mark to zero, and the picture area is initialised to the minimum rectangle containing current block in step **306**.

The process **317** is to extract the rectangle area of picture image and consists of two steps. Firstly, process **318** extracts the picture blocks. Then this area will further grow to its neighbour pixels in process **319** if necessary.

Switch **307** checks whether there is a neighbour block of current picture area whose mark is **1**. If the answer is **YES**, mark this block **0** in **308** and extend picture area to a new rectangle area containing this block in process **309**, go back to switch **307**. If the answer is **NO**, all neighbour blocks are not picture block. We finish process **318** and go to switch **310**.

Switch **310** checks that whether the rectangle picture area is big enough by comparing the length and width to the preset value (default $2W$). If answer is **NO**, there is no picture area found and turn to switch **304**. Otherwise, the answer is **YES**, we store the location information of the picture area in **311**.

Process **319** comprising following steps refines the picture area. Switch **312** checks whether there is a fore-pixel in the neighbour pixels of current area. If the answer is **YES**, process **313** extends the picture area to the new rectangle picture area containing the found fore-pixel and we go back switch **312**. If the answer is **NO**, all neighbour pixels of current picture area are back-pixels. Process **319** finishes and save this rectangle picture as a picture image area in process **314**. Process **315** appends this picture image area to the list of picture images then we go back to switch **304**.

Process **401** generate the style of characters

$\Omega=\{(i,j)\,|\,I(i,j)<P, i=0,1,\dots,h-1,\ j=0,1,\dots,w-1\}$ *where* $I(i,j)$ *is intensity of pixel at coodinates* (i,j), w *is width*, h *is height*. P *is estimated in step* 301

block distance of two pixels are defined as $d((i_1,j_1),(i_2,j_2)) = |i_2-i_1|+|j_2-j_1|$

$$d_{lt}=\min_{(i,j)\in\Omega}(d((0,0),(i,j)))\ ,\ d_{lb}=\min_{(i,j)\in\Omega}(d((0,h-1),(i,j)))$$

$$d_{rt}=\min_{(i,j)\in\Omega}(d((w-1,0),(i,j)))\ ,\ d_{rb}=\min_{(i,j)\in\Omega}(d((w-1,h-1),(i,j)))$$

The style of this character is $(w,h,d_{lt},d_{lb},d_{rt},d_{rb})$.

We define three sets L_0, L_1, L_2 for process **109**. L_0 is the collection of character images blocks. L_1 is the collection of the character code information of the character image blocks. L_2 is library of character templates used to save the images of character templates. Switch **402** checks whether L_0 is empty, if it answers **YES**, it means all character blocks have been processed, then data **403** comprising L_1 and L_2 will be outputted and terminate the process **109**. Otherwise, the answer is **NO** in step **402**, we will get the next character block T in L_0 in process **404**. **414** is process of matching character block T against templates in L_2. We start from the head of L_2. Check whether all templates in L_2 have been used in 406.

If the answer is YES, it means T is a new type of character, in step 407 we append it to L_2 as a new character template TL, save the code information of T against TL to L_1, and remove T from L_0, then go back to switch 402.

Otherwise, if the answer is NO in 406,get the character template TL from L_2 in step 408. We match T against TL by two steps, first match T against TL in process 409 by their style, Switch 410 checks the result of process 409, if the answer is NO, go to 406. If the answer is YES, then match T against TL by morphological character matching method in 411.

411 uses a morphological approach with which the matching of two characters is fast and accurate compared to the conventional matching method such as the matching by the grey scale similarity. The new measurement based on morphological is better than Euclidean distance measurement and Hausdorff measurement in the case of noise environment due to the stability of the measurement.

The new morphological operator measures the size of the difference image of two images (one is the template and the other is character block). Assume the two images are f and g, the difference image f-g is defined as follow

$$(f-g)(x,y)=\begin{cases}1, & F((x,y)_f)+F((x,y)_g)<4 \text{ and } |f(x,y)-g(x,y)|>C_M \\ 0, & \text{otherwise}\end{cases},$$

threshold $C_M = 32$.

The different image f-g is a binary image, in the other words, it is a binary set.

Define the size of set A of structure element B as $e(A)_B = \sup_{\alpha}\{A \circ \alpha B \neq \phi\}, \alpha \in \Re$

where $A \circ B$ *is normal morphological open operator.*

The new measurement of the difference between two binary sets can be defined as $S_B(f,g) = e(f-g)_B$, *where* B *is square structure element of size 1.*

The similarity measure of two sets f, g is $M(f,g) = \max\{S_B(f-g),\ S_B(g-f)\}$. The new measurement is symmetric in the sense of the distortion is concave distortion or convex distortion; however, the Hausdorff measurement is not symmetric [8].

If the measure is less than the average size of the noise region, the matching is success. We develop a fast algorithm based on this theory for matching of character problem. The measure of the difference is modified as $M(f,g) = S_B(f-g)$. For the matching for character image with resolution no less than 72, if the measure is less than 2, the matching of character against template is success. The algorithm is as following,

```
Algorithm $M_1$
    1. Suppose $(f - g)(x)$ is a sequence with length $m$. $x \leftarrow 0$,
    2. if $(f - g)(x) = 0$, go to step 5
    3. if $(f - g)(x+1) = 0$, $(f - g)(x) \leftarrow 0$, go to step 5
    4. $x \leftarrow x+1$
    5. if$(x < m-1)$ $x \leftarrow x+1$, go to step 2
    6. end
Algorithm $M_2$
    1. Suppose $(f - g)(x)$ is a sequence with length $m$. $x \leftarrow 0$,
    2. if $(f - g)(x) = 1$ and $(f - g)(x+1) = 1$, go to step 5
    3. if$(x < m-1)$ $x \leftarrow x+1$, go to step 2
    4. character matches against template, go to step 6
    5. character does not match against template, go to step 6
    6. end
```

The condition is weak or not depends on the structure element used in the algorithm M_1 and the associated part of algorithm M_2. Here the condition is strong means that it is difficult to match a character against template. On the contrary, the condition is weak means it is very easy to match a character against a template. Strong condition will decrease the compression ratio slightly but weak condition will generate false matching and the reconstructed character may not be correct when the scanned document image quality is very poor. The order of line, circle to square corresponds to the conditions from strong to weak.

Note:

1. If algorithm M_1 performs only along row direction, the structure element used in the matching algorithm is line of horizontal direction. This element is good enough for the English character matching.
2. If algorithm M_1 performs only along column, the structure element used in the matching algorithm is line of vertical direction.
3. If algorithm M_1 performs along both row and column directions, the structure element used in the matching algorithm is circle. Circle structure element works well for character of most languages.
4. If algorithm M_1 performs along row direction followed by column direction and then performs along column direction followed by row direction, the structure element used in the matching algorithm is square.
5. For the structure element of lines we only need apply algorithm M_2 along same direction as M_1 does. For the structure element circle algorithm M_2 performs at either horizontal direction or vertical direction. For the structure element square, algorithm M_2 performs at both horizontal and vertical directions before we can conclude that the match is success.

Switch 412 checks whether *T* matches against *TL*. If the answer is NO, go back to switch 406.Otherwise, the answer is YES, information of *T* is appended to L_1 and code of T is index of pattern *TL* in L_1, then process 413 removes *T* from L_0, then we go to switch 402.

Dec d r 20

Figure 2 is WDIC decoder 20 that is the reverse process of encoder 10. Decoder 20 starts from compressed bit stream 201 of document image. Process 202 separates 201 to three parts based on the formats of compressed document image described in 118. These three parts are compressed bit stream 203 of background image, compressed bit stream 206 of character image blocks and compressed bit stream 209 of picture image blocks.

Data 203 is decoded by wavelet based SAQ decoder 204 to generate the background image. Data 206 is decoded by character decoder 207 to generate the information of character codes of character image blocks and character template library. Data 209 is decoded by wavelet based SAQ decoder 210 to generate the picture image blocks 211. Data 205, 208 and 211 will be combined to document image 213 in process 212.

Claims

What is claimed:

1. The method of encoding of document image comprising of the steps of extracting two main categories of picture areas and character/line areas from the document image and,
   obtaining the residue image by subtracted these two categories areas from document images, and,
   classifying the character/line according to the templates dynamical generated, and
   encoding the residue image by wavelet based SAQ method.
2. The method of extracting picture areas from document image according to claim 1.
3. The method according to claim 2 comprising marking the blocks partitioned from document image based on features of their wavelet coefficients,
4. The method according to claim 2, further comprising hierarchical picture area extracting method comprising steps of
   extracting the picture blocks first to generate the initial picture area and,
   refining the picture area to cover the neighbour picture pixels of original area.
5. The method according to claim 1, further comprising the method of extracting character/line areas from document image,
6. The method according to claim 5 comprising of special definition of the connectivity,
7. The method according to claim 5, further comprising of the method of extracting the character/line areas,
8. The method according to claim 1, further comprising the method of classifying character/line,
9. The method according to claim8 comprising generating style of the character/line areas,
10. The method according to claim 8 further comprising the hierarchical matching of the character/line area against character/line templates comprising steps of matching the styles of character/line areas against styles of templates first and, matching of the character/line areas against template,
11. The method of matching of the character/line areas against template comprising of morphological matching,

12. The method according to claim 11, comprising specific character/line area matching algorithms M_1 and M_2,
13. The method according to claim11, further comprising the method of using different structure element for the different kind of document image,
14. The method according to claim 1, further comprising of bit plane storage of the compressed stream of the document image by the order of character/line, picture and background image which can be progressive decoding by associated decoder,
15. The associated decoder of encoder in claim 1.

References

1. G.Story,L.O'Gorman, D.Fox, L.Ghaper, and H.Jagadish. The Right Pages image- based electonic library for alerting and browsing. *IEEE Computer, 25(9):17-26* 1992
2. T.Phelps and Wilensky. Towards active, extensible, networked documents: Multivalent architecture and applications. *In proceedings of the 1st ACM International Conference on Digital Libraries*, pages 100-108,1996
3. I.H.Witten, A.Moffat, and T.C.Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, NewYork, 1994
4. P.Haffiner, L.Bottou, P.G.Howard, P.Simard, Y.Bengio, and Y.L.Cun Browsing through High Quality Document Images with DjVu. IEEE 1998
5. J.M.Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients", IEEE Trans. On Signal Processing, Vol. 41, No. 12, Dec, 1993, pp. 3445-3426
6. Said, and W. A. Pearlman, "A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees", IEEE Trans, on Circuits and Systems for Video Technology, Vol. 6, No. 3, June 1996, pp. 243-250
7. Gillbert Strang, Truong Nguyen, Wavelets and Filter Banks, Wellesley-Cambridge Press, 1996.
8. W. Gong, Q. Y. Shi, and M. D. Cheng, Shape and image matching by use of morphology, Proc. 11th Int. Conf. On Pattern Recognition, vol. 2, 673—676, The Hague, The Netherlands, 1992.

Encoder 10

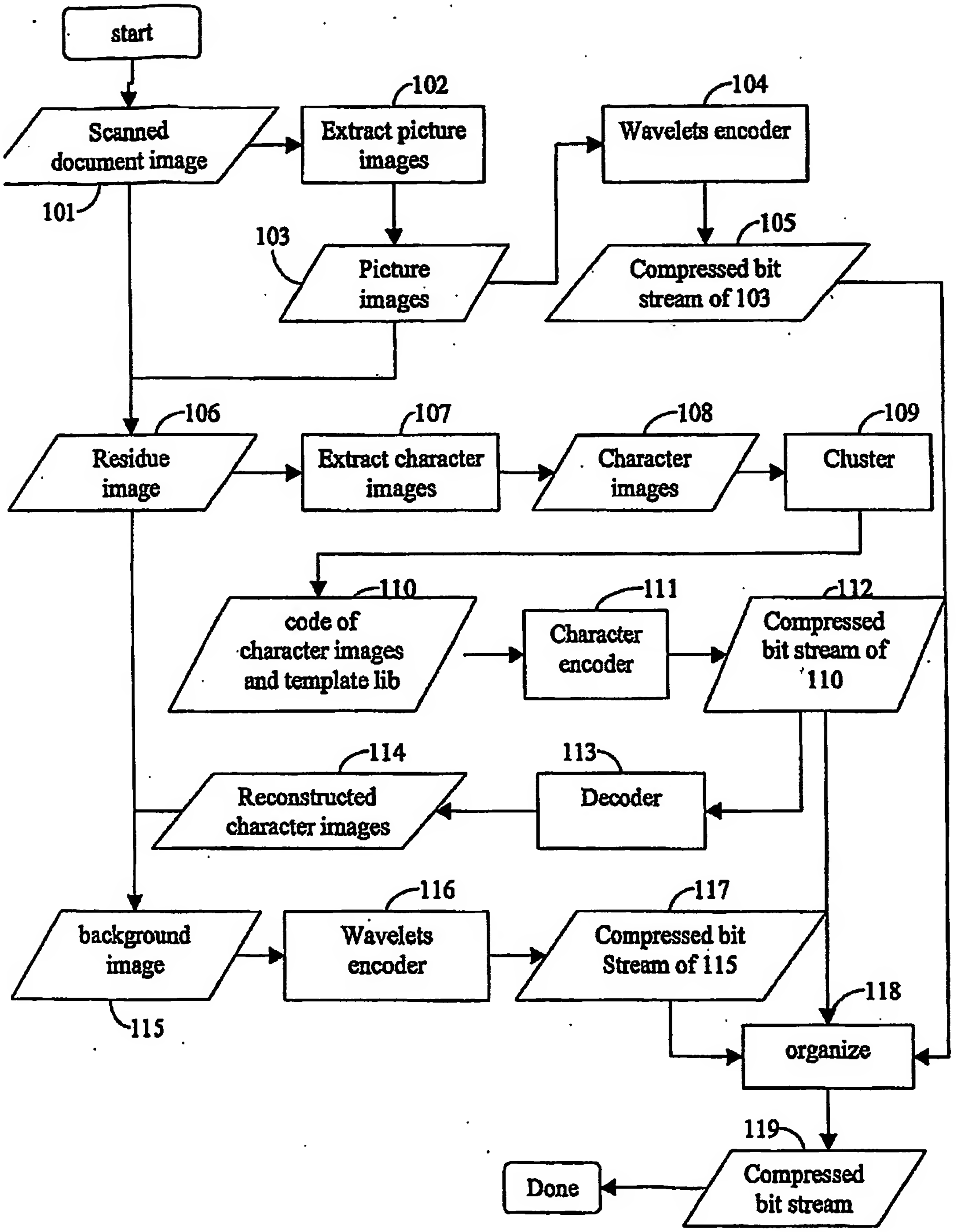


Figure 1

Decoder 20

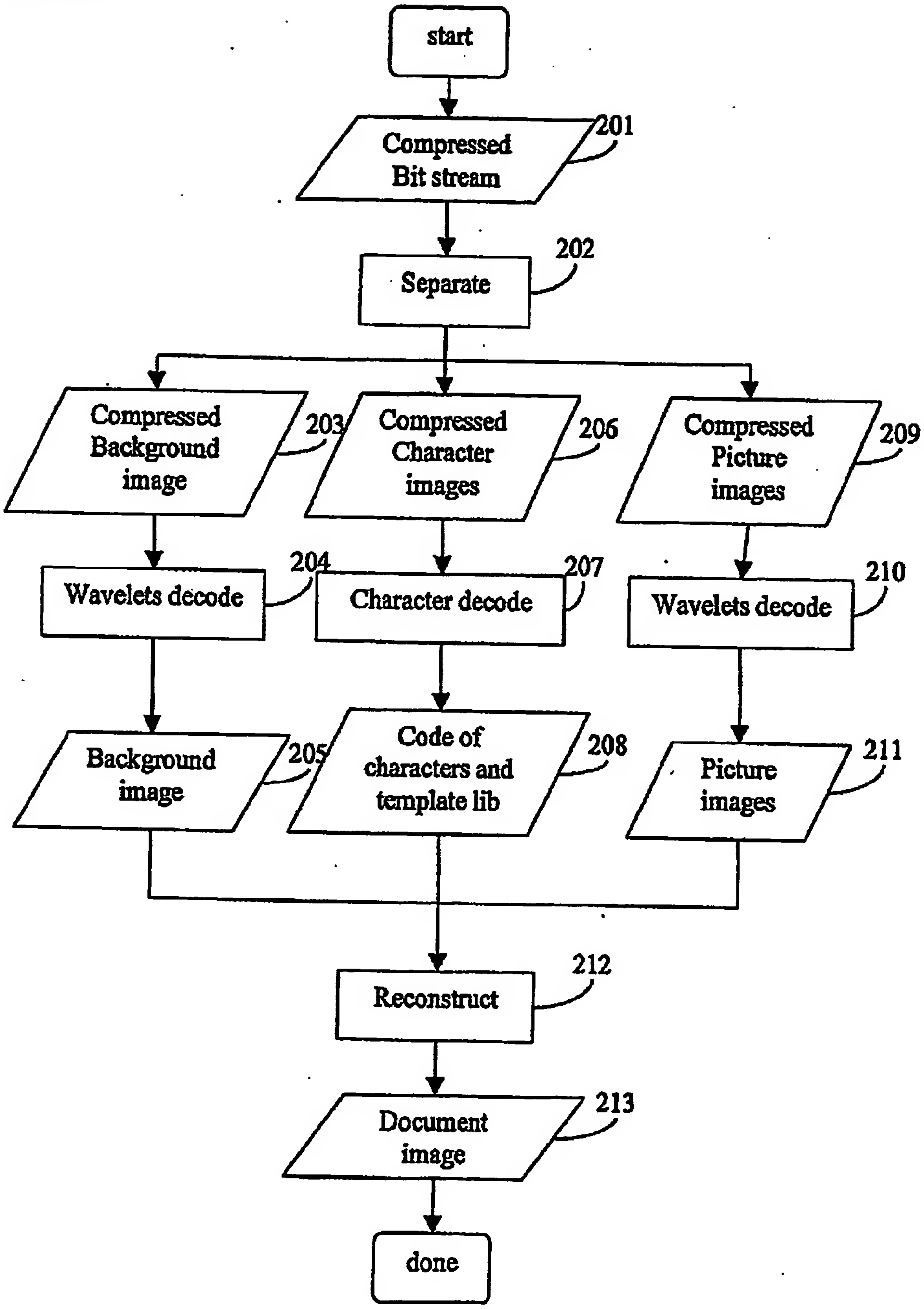
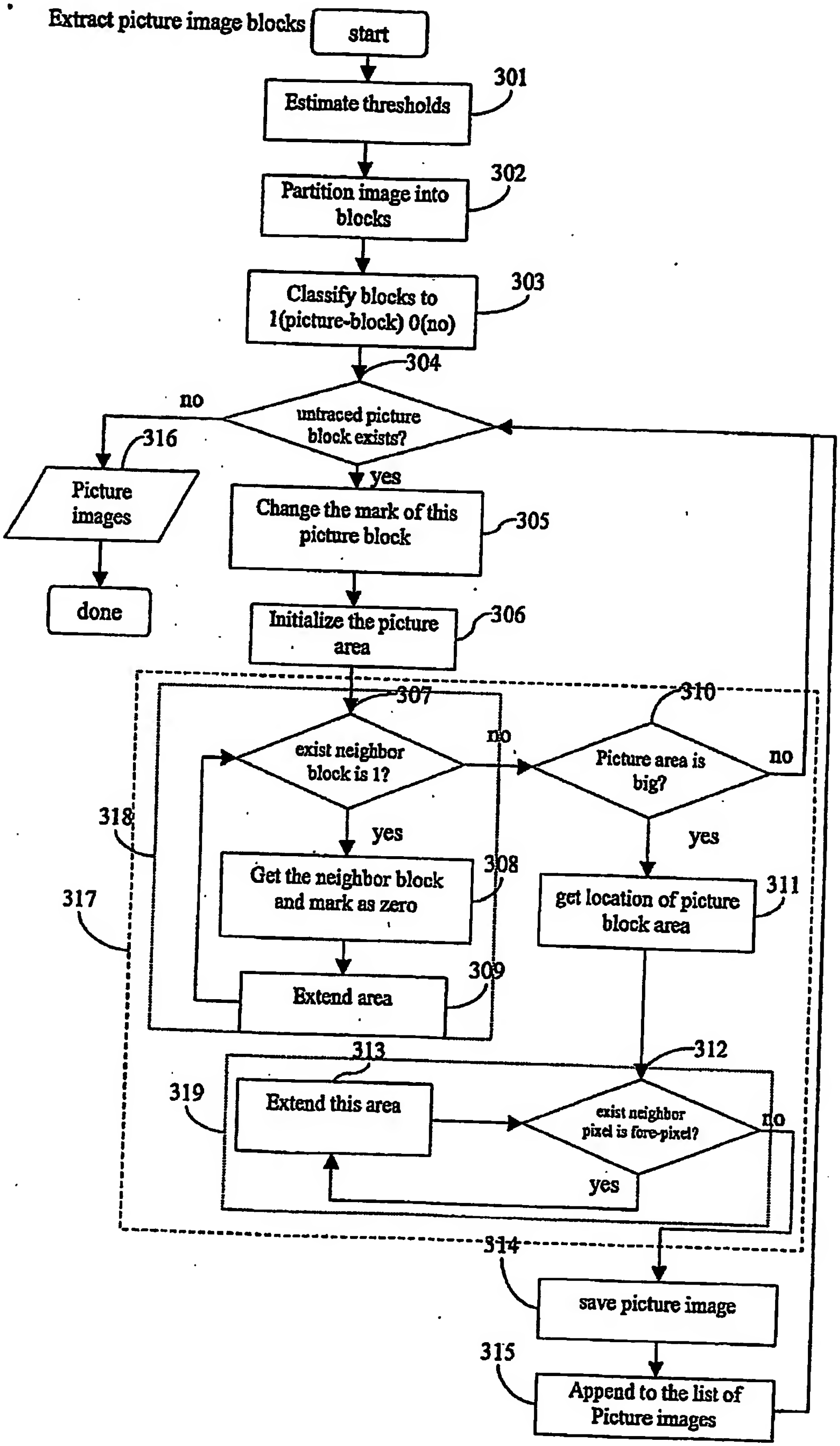
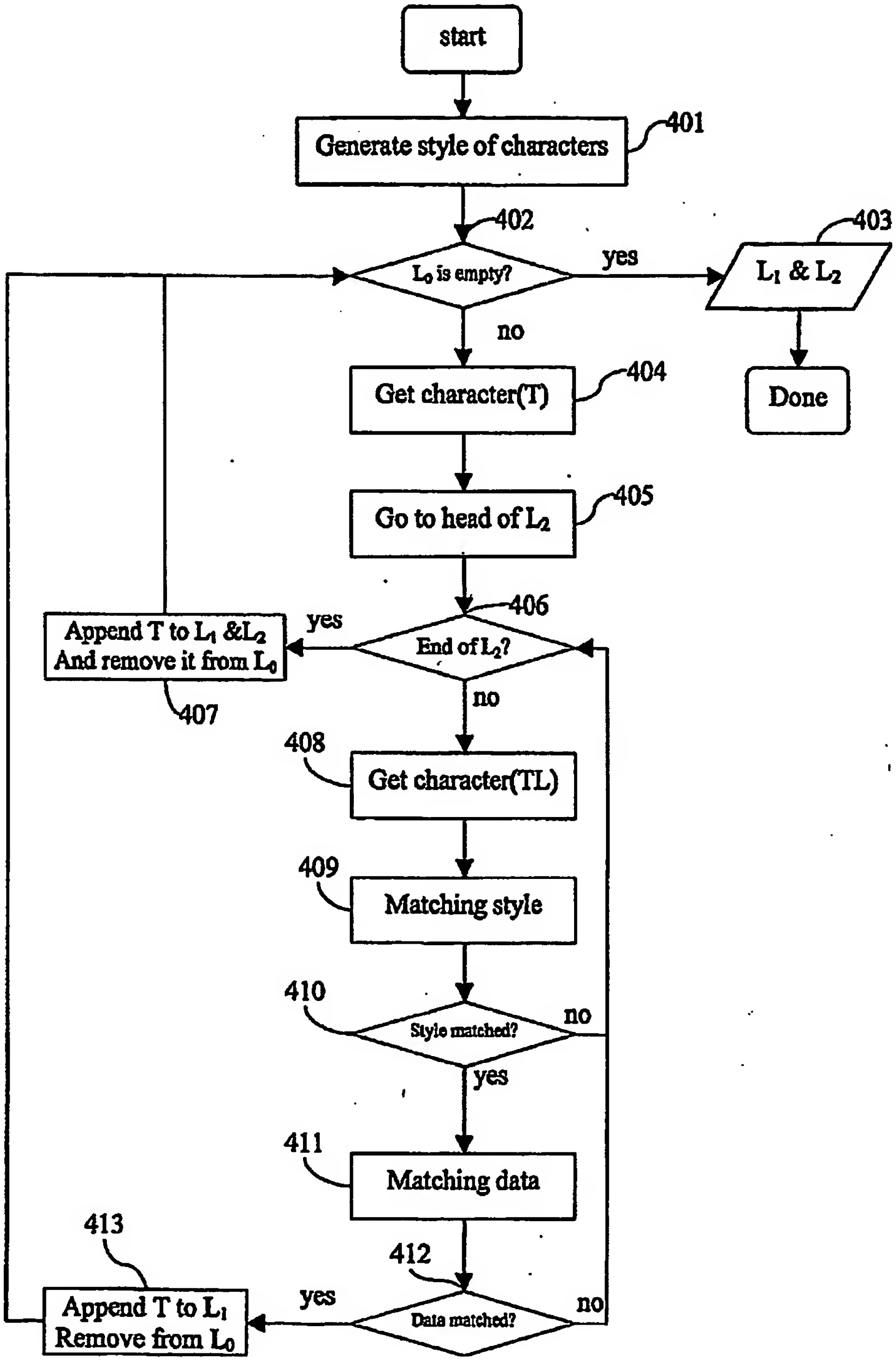


Figure 2


Extract picture image blocks
start
Estimate thresholds
301
Partition image into blocks
302
Classify blocks to 1(picture-block) 0(no)
303
304
untraced picture block exists?
no
yes
316
Picture images
done
Change the mark of this picture block
305
Initialize the picture area
306
307
exist neighbor block is 1?
no
yes
Get the neighbor block and mark as zero
308
Extend area
309
310
Picture area is big?
no
yes
get location of picture block area
311
318
317
313
312
319
Extend this area
exist neighbor pixel is fore-pixel?
no
yes
314
save picture image
315
Append to the list of Picture images

Clustering character images

L_0 is the list of character image blocks
L_1 is the list of code information of character image blocks
L_2 is library of character templates



start
Generate style of characters
401
402
L_0 is empty?
yes
403
L_1 & L_2
Done
no
Get character(T)
404
Go to head of L_2
405
406
End of L_2?
yes
Append T to L_1 &L_2 And remove it from L_0
407
no
408
Get character(TL)
409
Matching style
410
Style matched?
no
yes
411
Matching data
412
413
Append T to L_1 Remove from L_0
yes
Data matched?
no